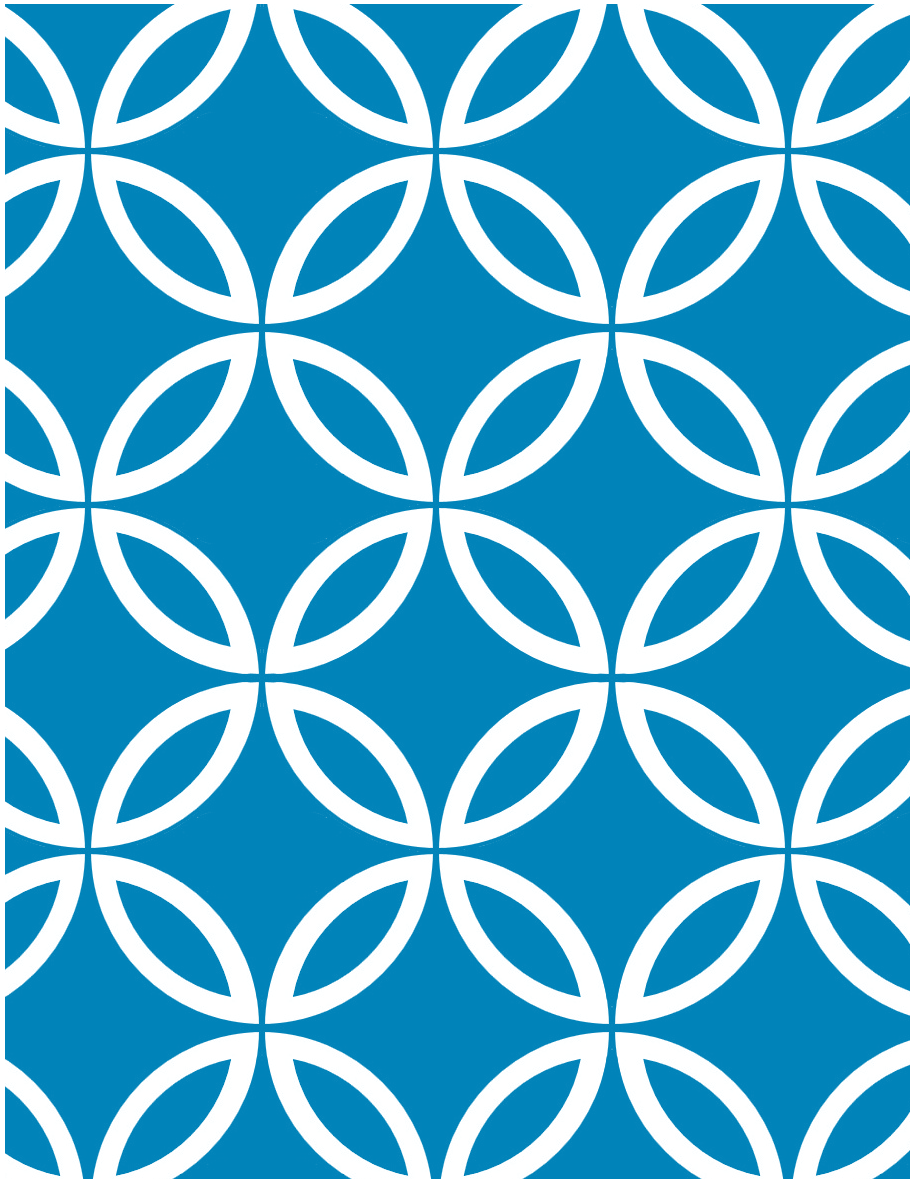


CONSIDERATIONS FOR POWER IN META-ANALYSIS

Terri Pigott, Associate Provost for Research
Loyola University Chicago, USA

OBJECTIVES – A TEACHING SESSION

Understand how power analysis can help in planning for a meta-analysis

Outline the steps for conducting power analysis in two scenarios

- The random effects mean effect size
- Subgroup analysis

Discuss the challenges for power in meta-analysis such as in meta-regression

WHY CONDUCT POWER FOR MA?

When planning a primary study, researchers need to know how many participants are needed to detect the expected experimental difference between groups

But in a meta-analysis, researchers have no control over our sample size, i.e., the number of eligible studies for a review

So, why do a power analysis before even starting the systematic review?

REASONS TO CONDUCT POWER FOR MA

Though we don't know how many eligible studies will be identified, we can plan for the magnitude of the effect size that can be detected in the analysis

Reporting the potential power of the meta-analysis in the protocol (or grant application) increases the transparency of the systematic review and meta-analysis

Knowing the potential power helps in planning analyses and interpreting results – what meta-analysis models might be possible with a sufficient number of studies?

TO PREPARE TO CONDUCT A POWER ANALYSIS

Conduct an informal scoping review

- How many studies are potentially available for the systematic review and meta-analysis?

Understand the nature of the literature under review

- How are studies typically conducted in this area?
- How big is the typical sample size for the studies?

Decide on the value of a clinically important effect size

- What is a clinically important difference between the treatment and control group?
- What is an important correlation in this context?

TO ILLUSTRATE POWER ANALYSIS

We will start with conducting power analysis for the overall mean effect size in a random effects model

Let's imagine that you are planning a systematic review and meta-analysis, and your first task is to estimate the random effects overall mean effect size

Your first question will be: What is the power of my meta-analysis for finding a random effects mean effect size of a particular value, θ?

Recall that we compute power for a particular statistical test. Here we will focus on the significance test for the random effects mean effect size, or the Z-test

TO COMPUTE POWER IN META-ANALYSIS

We need to make a set of assumptions, just as we do in planning a primary study

These assumptions are related to

- The number of studies that will be eligible for the meta-analysis
- The “typical” sample size of these studies
- The effect size of substantive interest and its variance
- The amount of heterogeneity among studies (in our random effects model)

ASSUMPTIONS FOR POWER ANALYSIS FOR META-ANALYSIS (RE MODELS)

1. Establish a critical value for statistical significance, c_α
2. Decide on a value or range of values for substantively important effect sizes, θ
3. Estimate the number of eligible studies, k
4. Estimate the within-study sample sizes to compute the "typical" within-study effect size variance, v
5. Estimate the variance component, τ^2
6. Compute power for a given test in random effects meta-analysis

DISCUSSION STEPS IN THIS SECTION

We will discuss power analysis in general and how we compute it

We will discuss how we actually think about the assumptions we need to make in a power analysis

POWER OF *Z*-TEST IN WORDS

We use the Z-test for the statistical significance of our obtained effect size – to see if our effect size is significantly different from 0

We want to know if our meta-analysis has the power to detect a value of the mean effect size that is different from zero

In essence, we are trying to figure out the power of the test to reject the null hypothesis of effect size equal to 0 (no matter what kind of effect size)

The power of the test will depend on our alternative hypothesis – whether we think the effect size is bigger or smaller than 0 (one-tailed) or just different from zero (non-directional)

Z- TEST FOR THE RANDOM EFFECTS MEAN

$$Z = \frac{\overline{T}_\bullet - 0}{\sqrt{v_\bullet}}$$

where $\overline{T}_\bullet$ is the mean effect size of interest and $v_\bullet$ is the random effects variance for the mean effect size equal to

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} 1/(v_i + \widehat{\tau^2})}$$

where v_i is the sampling variance for the effect size in the *i*th study, and $\widehat{\tau^2}$ is the estimated variance component

WHEN OUR EFFECT SIZE IS EQUAL TO 0

When the null hypothesis is true (when our effect size is equal to 0) or in notation

Null hypothesis: H_0: $\theta = 0$

Z has a standard normal distribution with mean equal to 0 and variance equal to 1

BUT OF COURSE WE DON'T THINK EFFECT SIZE IS 0

We usually think either of these two alternatives:

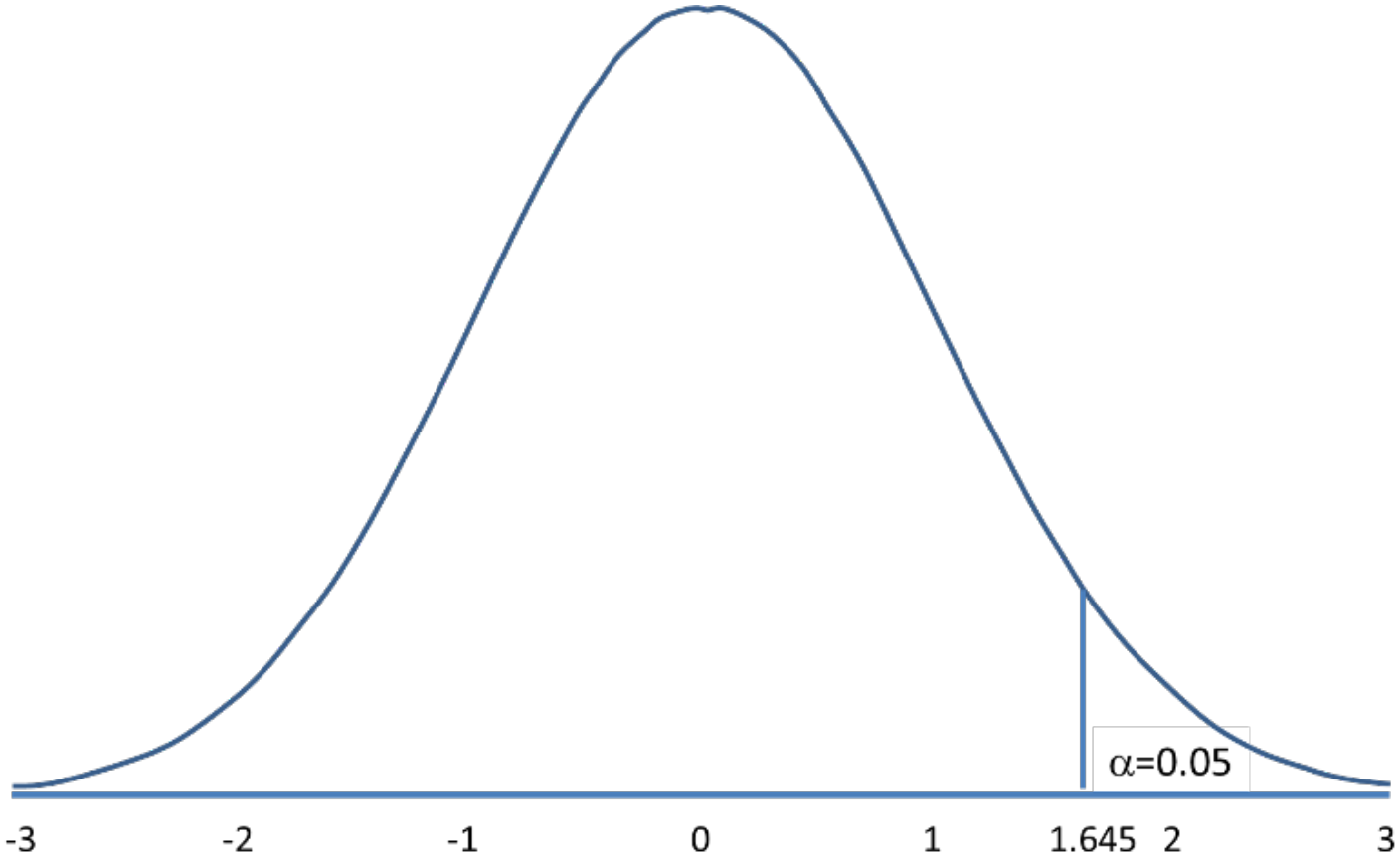
One-sided test: H_a: $\theta \geq 0$ or H_a: $\theta \leq 0$

Two-sided test: H_a: $\theta \neq 0$

LET'S THINK ABOUT THE ONE-SIDED TEST

We will reject the null hypothesis H_0 if the value of Z is greater than the critical value, c_α of the standard normal distribution

When $\alpha = 0.05$, the critical value for the standard normal distribution is 1.645



WHEN THE NULL HYPOTHESIS IS FALSE

And we think H_a: $\theta \geq 0$:

Z has a normal distribution with a variance of 1 and a mean equal to

$$\lambda = \frac{\theta - 0}{\sqrt{v_\bullet}}$$

where θ is the target value of the effect size

POWER IN WORDS

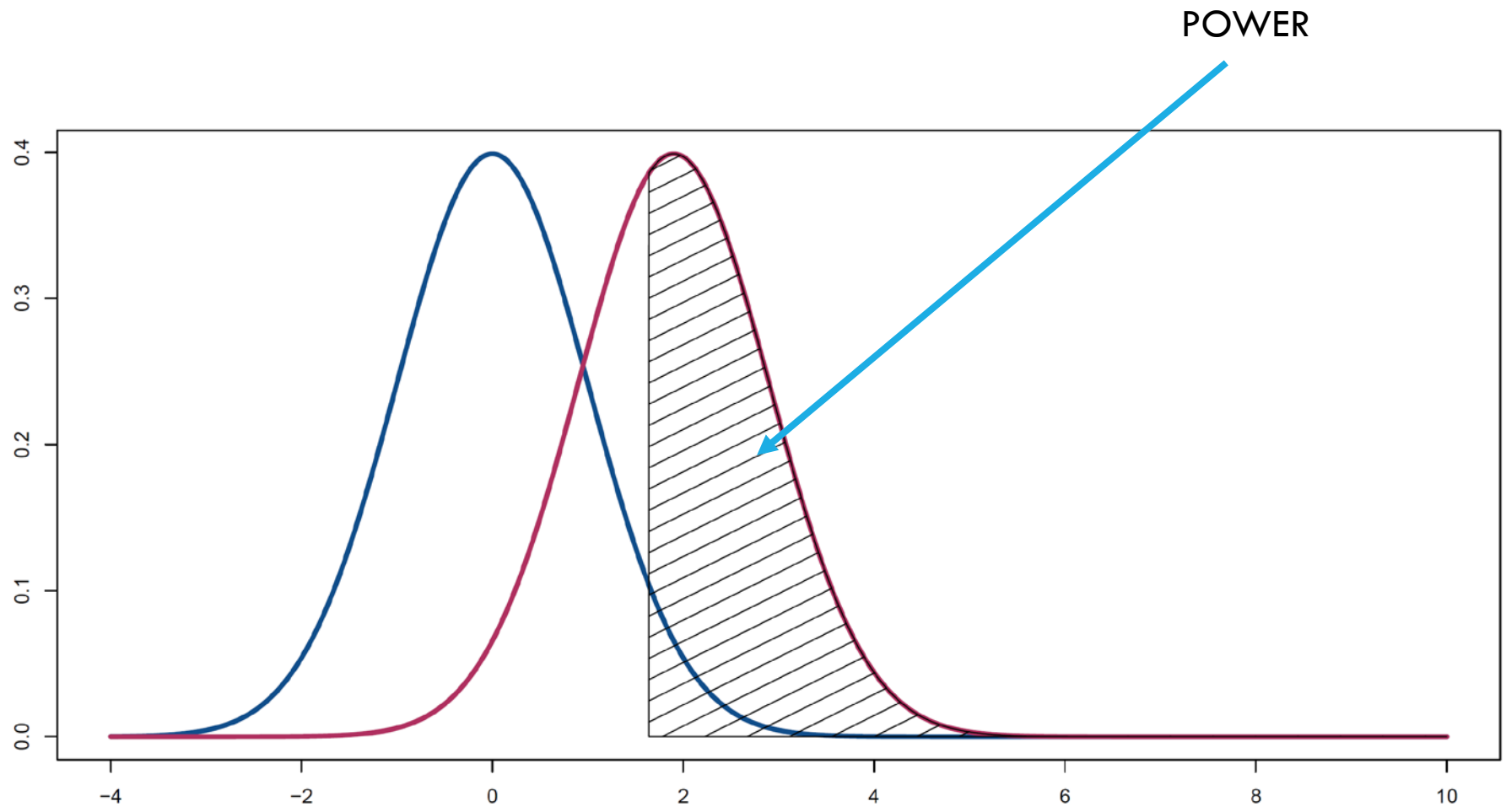
We now have two different normal distributions we are comparing:

- The standard normal distribution under the null hypothesis, and
- The normal distribution with mean θ under the alternative hypothesis H_a

We want to know the proportion of the normal distribution under the alternative distribution that exceeds the critical value, c_α, in the standard normal distribution

Blue is standard normal, distribution under null hypothesis

Red is distribution under the alternative hypothesis where $\theta = 2.5$



WE COMPUTE POWER BY

$$p = 1 - \Phi(c_\alpha - \lambda)$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution, or the area under the standard normal curve from $-\infty$ to x.

IN OUR EXAMPLE

Cross-hatched area is area under normal curve for H_a that exceeds value of $c_\alpha - \lambda$ when $c_\alpha = 1.64$ and $\lambda = 2.5$

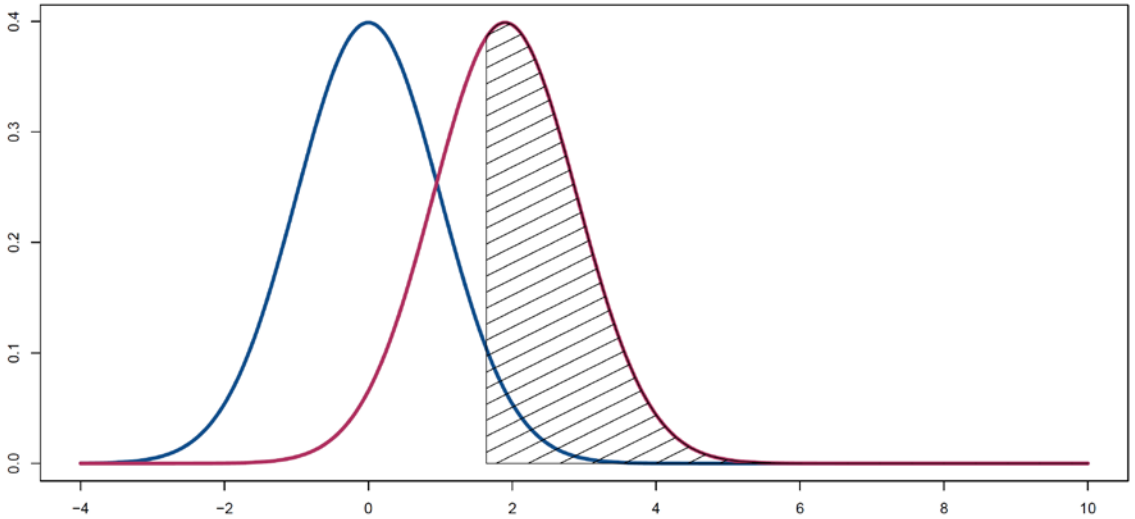
Exact value:

$$p = 1 - \Phi(1.64 - 2.5)$$

$$= 1 - \Phi(-0.86)$$

$$= 1 - 0.19 = 0.81$$

Power under H_a is 0.81



TWO-TAILED TEST

When H_a: $\theta \neq 0$

Power is given by

$$p = 1 - [\Phi(c_{.5\alpha} - \lambda) - \Phi(-c_{.5\alpha} - \lambda)]$$

$$p = 1 - \Phi(c_{.5\alpha} - \lambda) + \Phi(-c_{.5\alpha} - \lambda)$$

SO NOW WE KNOW HOW TO DO THIS CONCEPTUALLY

How do we actually do this?

WE NEED TO GET THESE QUANTITIES FOR POWER

$$\lambda = \frac{\theta - 0}{\sqrt{v_\bullet}}$$

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} 1/(v_i + \tau^2)}$$

ASSUMPTIONS NEEDED FOR POWER

The critical value of the test, c_α

The substantively important values for the effect sizes, θ

The number of eligible studies, k

The typical sample size within studies, N, so that we can get the sample variance for the "typical" effect size, v

The amount of between-study variance, τ^2

CAVEAT

We are assuming in this section that we only have **one** effect size per study

You will see or already know that this is never the case – studies provide multiple effect sizes per study

Thus, our power analysis will be approximate only

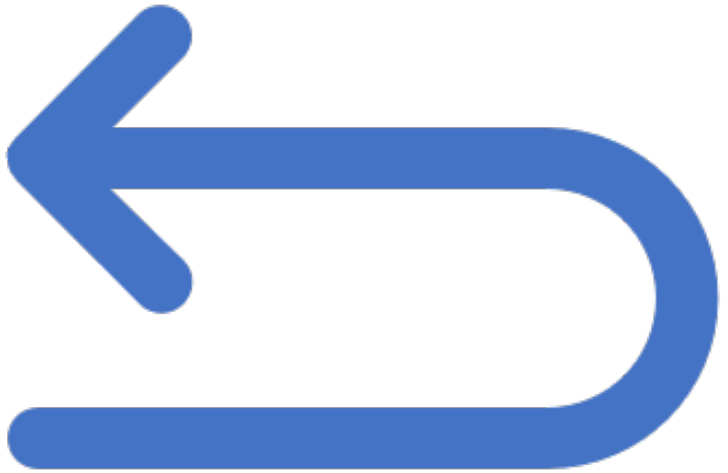
BACK TO OUR ASSUMPTIONS

We will go through each one of them to discuss how we might generate ideas for the power analysis

ASSUMPTION 1: c_α

First one is easy – what significance (α) level?

Usually $\alpha = 0.05$



WHAT IS A SUBSTANTIVELY IMPORTANT EFFECT SIZE?

Good question…

It depends on the context and the measures used in the studies you will review

My old example was for the college entrance exam test in the US (mostly because I had teens in the house at that time)

What effect size would I consider important before I paid hundreds of dollars for test prep?

10 points? 20 points? 50 points?

THINKING ABOUT SUBSTANTIVE EFFECT SIZE FOR POWER ANALYSIS

This issue depends on the context and scale of measurement typical for the set of studies being reviewed

Given the difficulty in interpreting effect sizes in general, I suggest starting with a concrete example from your set of potential studies

For example:

- What is a policy-relevant difference between treatment and control groups in terms of a given measurement scale or percentile gain?
- What change or difference in number of successes is important?
- What constitutes a substantively important correlation?

EXAMPLE: HS SCIENCE INTERVENTION

Let's say that ACT Science scores are eligible measures for a high school science intervention

In 2016, the average ACT Science score for the US was 20.8 with a standard deviation of 5.6

An effect size of 1.0 would mean a difference in the treatment and control conditions of 5.6 points

An effect size of 0.5 would translate into a difference of 2.8 points

Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. (2007). Empirical benchmarks for interpreting effect sizes in research. MDRC Working Papers on Rsearch Methodology.

https://www.mdrc.org/sites/default/files/full_84.pdf

ANOTHER SOURCE OF EFFECT SIZE INFORMATION

In this article, we argue that effect sizes should instead be interpreted with respect to empirical benchmarks that are relevant to the intervention, target population, and outcome measure being considered. We illustrate this point with three types of benchmarks: (1) normative expectations for change, (2) policy-relevant performance gaps, and (3) effect size results from similar studies. Our analysis draws from a larger ongoing research project that is examining the calculation, interpretation, and uses of effect sizes measures in education research.[2] Thus we illustrate each benchmark with educational examples. The more general message — that effect sizes should be interpreted using relevant empirical benchmarks — is applicable to any policy or program area, however.

HILL ET AL. (2007) QUOTE

Effect Size Measures

Table 4

Summary of Effect Sizes from Randomized Studies

Achievement Measure	Number of Effect Size Estimates	Mean Effect Size	Standard Deviation
Elementary schools	389	0.33	0.48
Standardized test (broad)	21	0.07	0.32
Standardized test (narrow)	181	0.23	0.35
Specialized topic/test	180	0.44	0.49
Middle schools	36	0.51	0.49
High schools	43	0.27	0.33

SOURCES: Compiled by the authors from 61 existing research reports and publications (reporting on 95 independent subject samples).

NOTE: Unweighted means are shown across all effect sizes and samples in each category.

ASSUMPTION 2: SIZE OF EFFECT

Let's pick a range of θ's that correspond to potential effect sizes

Use substantive knowledge of the area of the meta-analysis or Hill et al.'s empirical values

ASSUMPTION 3: HOW MANY STUDIES, *k*?

We can use an informal scoping review to make educated guesses about the potential range of numbers of eligible studies

We could also argue for the importance of examining power at the lower expected bound of number of eligible studies

ASSUMPTION 4: SAMPLE SIZE WITHIN STUDIES, *N*

This is a much more difficult guess – how big is the "typical" study we expect in the systematic review?

Some questions to think about:

Is this an area with many RCTs?

Is this an intervention difficult to implement and likely studied with smaller samples?

ASSUMPTION 4: WITHIN-STUDY SAMPLE SIZE

We will assume that all studies have the same within-study sample size for now

We will use the "typical" sample size to compute the variance of the within-study effect size

The "typical" within-study effect size variance is needed to compute the variance of our target overall mean effect size

WE NEED TO GET THESE QUANTITIES FOR POWER

$$\lambda = \frac{\theta - 0}{\sqrt{v_\bullet}}$$

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} 1/(v+\tau^2)}$$

EFFECT SIZE VARIANCES

SMD: $v_i = \frac{n_E+n_C}{n_E n_C} + \frac{\theta^2}{2(n_E+n_C)}$

Fisher's Z: $v_i = \frac{1}{n-3}$

Log odds ratio: $v_i = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$

ASSUMPTION 5: THE VARIANCE COMPONENT?

How do we think about τ^2 ?

Consult prior meta-analyses

Think about the breadth of the research question for the review

Be guided by understanding of the substantive area

Use conventions for I^2 from the Cochrane Handbook (next slide)

QUOTE FROM COCHRANE HANDBOOK

Thresholds for the interpretation of I^2 can be misleading, since the importance of inconsistency depends on several factors. A rough guide to interpretation is as follows:

0% to 40%: might not be important;

30% to 60%: may represent moderate heterogeneity*;

50% to 90%: may represent substantial heterogeneity*;

75% to 100%: considerable heterogeneity*.

http://handbook-5-1.cochrane.org/ From section 9.5.2

RELATIONSHIP BETWEEN I^2 AND VARIANCE COMPONENT

Conceptually, I^2 is the percent of variation in the effect size that is attributed to between-studies differences

We can write I^2 as:

$$I^2 = \frac{\tau^2}{\tau^2 + v}$$

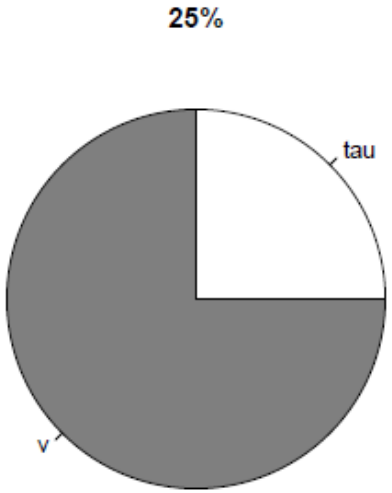
where v is a "typical" value of the within-study effect size variance

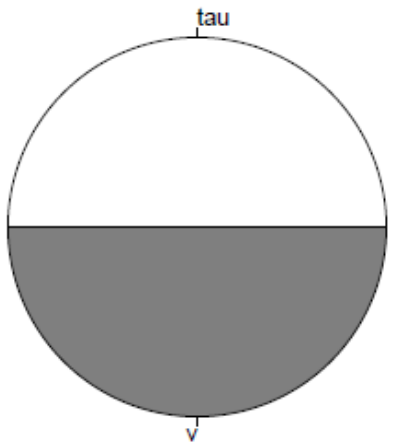
SO:

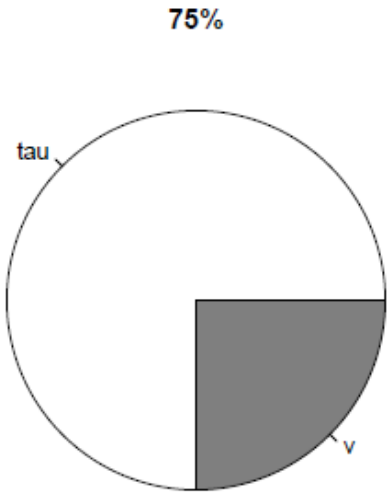
If $I^2 = 25\%$, then $\tau^2 = (v)/3$

If $I^2 = 50\%$, then $\tau^2 = 1.0(v)$

If $I^2 = 75\%$, then $\tau^2 = 3.0(v)$


25%
tau
v


50%
tau
v


75%
tau
v

LET'S PUT THIS TOGETHER IN AN EXAMPLE

Let's say we assume k=20 studies that have N=40 participants and we are interested in a standardized mean difference effect size of θ = 0.25

Let's also assume equal experimental and treatment sample sizes of 20

First we need to get the typical within-study variance, v, for our effect size θ

$$v = \frac{n_E+n_C}{n_E n_C} + \frac{\theta^2}{2(n_E+n_C)} = \frac{20+20}{20*20} + \frac{0.25^2}{2(20+20)} = 0.1008$$

We have $v = 0.1008$

For low heterogeneity: $\tau^2 = 0.1008/3 = 0.033$

For moderate heterogeneity: $\tau^2 = 0.1008$

For high heterogeneity: $\tau^2 = 3 * 0.1008 = 0.302$

LET'S GET VALUES FOR VARIANCE COMPONENT

NOW LET'S GET $v_\bullet$ FOR DIFFERENT TAU^2

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} 1/(v+\tau^2)} = \frac{1}{k/(v+\tau^2)} = \frac{v+\tau^2}{k}$$

with $k = 20$

Low: $v_\bullet = \frac{0.1008+0.033}{20} = 0.0067$

Moderate: $v_\bullet = \frac{0.1008+0.1008}{20} = 0.0101$

High: $v_\bullet = \frac{0.1008+0.302}{20} = 0.0201$

NOW WE CAN GET VALUES FOR LAMBDA

$$\lambda = \frac{\theta - 0}{\sqrt{v_\bullet}}$$

Low heterogeneity: $\lambda = \frac{0.25 - 0}{\sqrt{0.0067}} = 3.05$

Moderate heterogeneity: $\lambda = \frac{0.25 - 0}{\sqrt{0.0101}} = 2.49$

Low heterogeneity: $\lambda = \frac{0.25 - 0}{\sqrt{0.0201}} = 1.76$

POWER FOR 3 LEVELS OF HETEROGENEITY

One-tailed test of H_a: $\theta \geq 0.25$

Low: $p = 1 - \Phi(1.645 - 3.05) = 1 - \Phi(-1.405) = 0.92$

Moderate: $p = 1 - \Phi(1.645 - 2.49) = 1 - \Phi(-0.845) = 0.80$

High: $p = 1 - \Phi(1.645 - 1.76) = 1 - \Phi(-0.115) = 0.54$

SO....

When we have $k = 20$ studies, each with a total sample size of $N = 40$ participants:

We have power above 0.8 with low and moderate levels of heterogeneity to detect an effect size of 0.25

BUT with high levels of heterogeneity, we have power of only 0.54 to detect an effect size of 0.25

PRESENTATION OF POWER ANALYSIS

The next slide shows power for a range of assumptions for a standardized mean difference with the following assumptions:

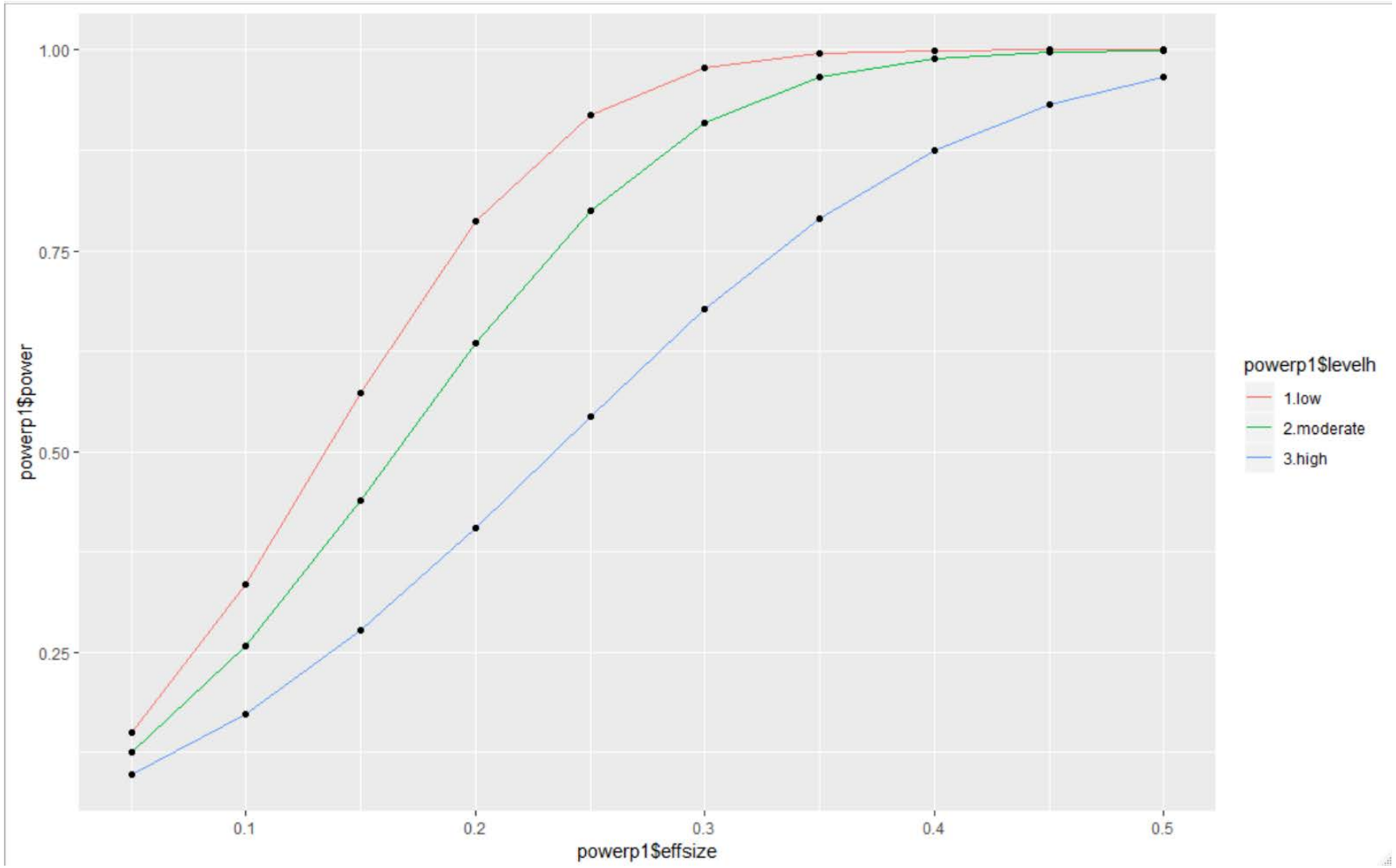
- Random effects mean effect size from 0.05 to 0.5
- Number of studies: k =20
- Sample size within studies: nTreatment = nControl = 20
- Three levels of heterogeneity: low, moderate and high as in our prior slides

We could produce graphs with other assumptions such as varying the number of studies or sample size within studies

R CODE AVAILABLE ON PSYCHARCHIVES

http://dx.doi.org/10.23668/psycharchives.2451

R code for both graphs in the presentation


1.00
0.75
0.50
0.25
powerp1$power
0.1
0.2
0.3
0.4
0.5
powerp1$effsize
powerp1$levelh
1.low
2.moderate
3.high

POWER FOR OTHER META-ANALYSIS TESTS

We can conduct power for other tests in meta-analysis, but many of these computations will require making more assumptions than we have for the random effects mean effect size

- Hedges and Pigott (2001) discusses power for the mean effect size and tests of heterogeneity under both the fixed and random effects models
- Hedges and Pigott (2004) discusses power for moderators in meta-analysis

The next slides briefly illustrate subgroup analysis and meta-regression

EXAMPLE OF TEST OF SUBGROUP DIFFERENCE

Imagine that we are interested in whether the treatment effect differs between women and men

We have some studies in our sample that provide an effect size for women and an effect size for men

Our statistical test of interest is called the between-group test of homogeneity – are the mean effect sizes from each group the same?

We will need values for the clinically important difference between the groups, and the estimated number of studies that provide the effect size for women and men

ASSUMPTIONS FOR SUBGROUP DIFFERENCE (FIXED EFFECTS)

1. Establish a critical value for statistical significance, c_α
2. Decide on a value or range of values for substantively important difference between the groups, θ
3. Estimate the number of eligible studies that provide effect sizes for each group, m
4. Estimate the within-study sample sizes to compute the within-study effect size variance, v

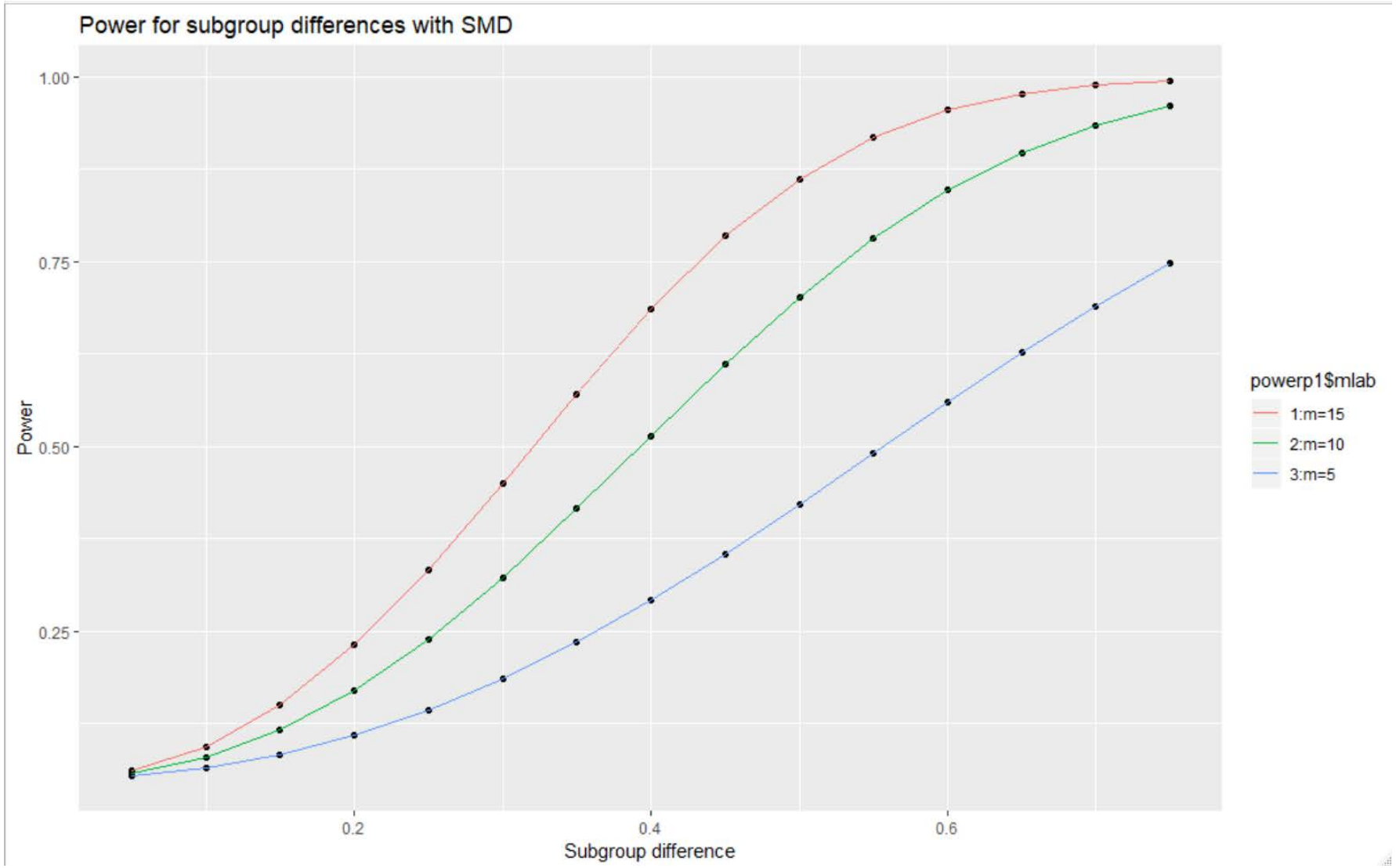
POWER CURVES FOR SUBGROUP DIFFERENCE

Next slide provides the power curves for the subgroup difference for the following scenario

- Equal numbers of studies for the two groups, *m*
- Range of effect size from 0.05 to 0.75
- Total within-study sample size of 20 (10 per experimental and control group)
- Fixed effects analysis

Note that imbalance in the groups will also impact power – the more imbalanced the groups, the lower the power

Random effects models will also have lower power


Power for subgroup differences with SMD
1.00
0.75
0.50
0.25
Power
0.2
0.4
0.6
Subgroup difference
powerp1$mlab
1:m=15
2:m=10
3:m=5

POWER FOR META-REGRESSION

While we can work out power computations for meta-regression, there are many assumptions that we need to make

For example:

- We need to guess at the values of the regression coefficients for each of the moderators included
- We need to know the balance of the covariates across the sample of studies

These assumptions are impossible to know prior to a meta-analysis

SUMMARY

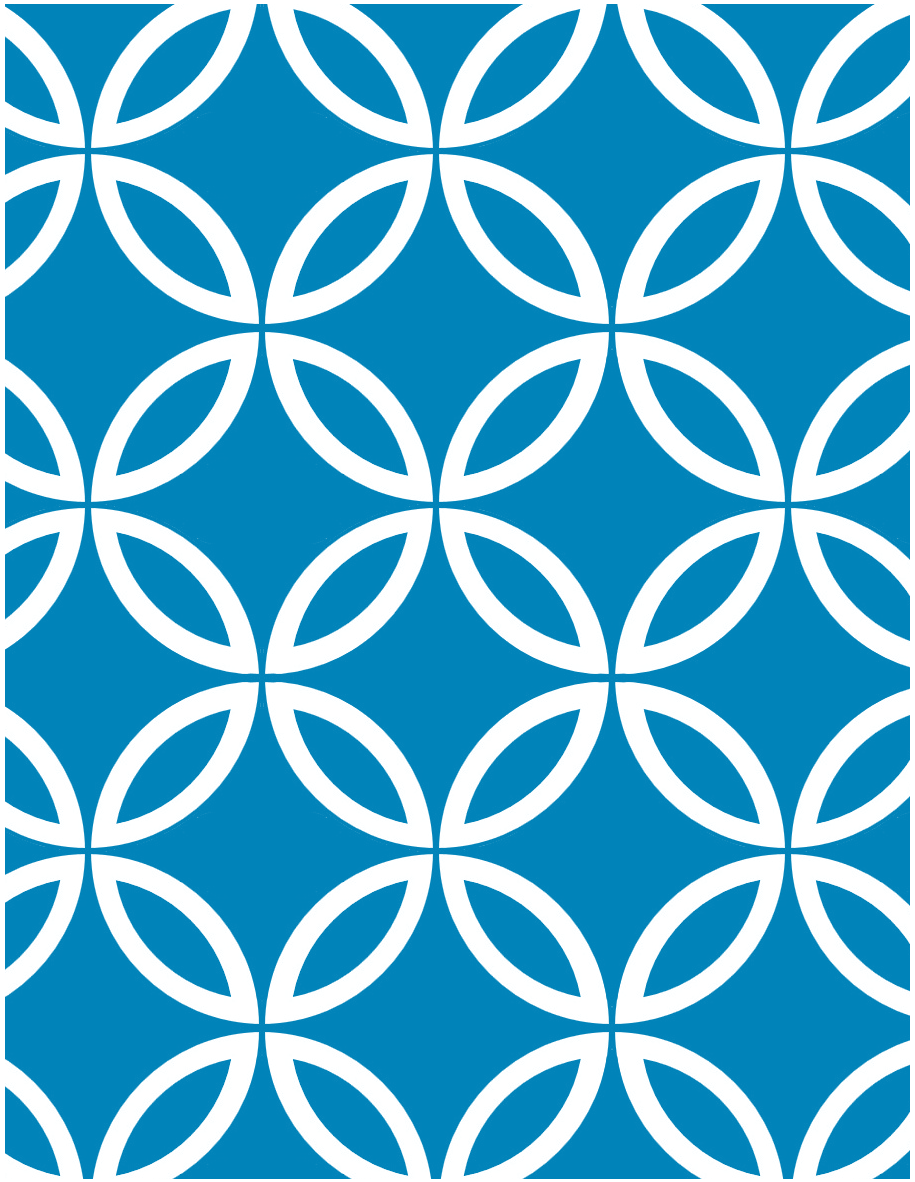
Power computations for meta-analysis can be useful for planning

Understanding the potential limitations of a meta-analysis *a prioiri* can guard against conducting too many analyses if the sample of studies is insufficient

Producing power computations under different assumptions can help in planning the most important analyses *a priori*

Future directions for research include:

- Power with multiple effect sizes per study in a multi-level framework
- Ways to think about power for meta-regression – what might be guidelines for exploring potential power for these models?



THANK YOU

Contact information: terri.pigott@gmail.com

REFERENCES

Pigott, T. D. (2012). *Advances in meta-analysis*. New York, NY: Springer-Verlag.

Hedges, L. V. & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426-445.

Hedges, L. V. & Pigott, T. D. (2001). Power analysis in meta-analysis. *Psychological Methods, 6,* 203-217.